Int. J. on Recent Trends in Engineering and Technology, Vol. 8, No. 1, Jan 2013

Performance Analysis of Hybrid Approach for Privacy Preserving in Data Mining

Savita Lohiya[1], Lata Ragha[2]
[1] SIES GST, Information Technology/ Navi Mumbai, India
Email: sgl.siesgst@gmail.com
[2] Terna Engineering College, Computer Department/ Navi Mumbai, India
Email: lata.ragha@gmail.com

***Abstract*— Now-a day's data sharing between two organizations is common in many application areas like business planning or marketing. When data are to be shared between parties, there could be some sensitive data which should not be disclosed to the other parties. Also medical records are more sensitive so, privacy protection is taken more seriously. As required by the Health Insurance Portability and Accountability Act (HIPAA), it is necessary to protect the privacy of patients and ensure the security of the medical data. To address this problem, released datasets must be modified unavoidably. We propose a method called Hybrid approach for privacy preserving and implemented it. First we randomized the original data. Then we have applied generalization on randomized or modified data. This technique protect private data with better accuracy, also it can reconstruct original data and provide data with no information loss, makes usability of data.**

***Index Terms*— Privacy preserving; Data Mining; Sensitive Data; k-anonymity; quasi-identifier**

I. Introduction

Privacy is an important issue when one wants to make use of data that involve individual sensitive information. As data mining usage are increased,large volumes of personal data are regularly collected and analyzed. Such data include shopping habits, medical history, criminal records, credit records etc. On the other hand, such data is an important asset to business organizations and governments both to decision making processes and to provide social benefit, such as medical research, crime reduction, national security, etc. This is mainly concerned with data custodians such as hospitals, government agencies, insurance companies, and other businesses that have data they would like to release to analysts, researchers, and anyone else who wants to use the data. The overall intent is for the data to be used for the public good. With the rapid development of internet technology, privacy preserving data publication has become one of the most important research topics and become a serious concern in publication of personal data in recent years. However, for data owners who are becoming increasingly concerned about their privacy due to the data which contains some personal information about individuals has been published by government departments and some business agencies, such as health insurance companies, hospitals. With this increasing, new threats to privacy of the. individual are also increases.Thus, an interesting new direction of data mining research has been developed, known as privacy preserving data mining (PPDM). The aim of these algorithms is the extraction of relevant knowledge from large collection of data, while protecting private information simultaneously. Privacy preserving means to prevent information disclosure due to legitimate access to the data. Thus, privacy preserving is different with conventional data security, access control and encryption technology that tries to prevent information disclosure against illegitimate means . The main consideration in privacy preserving data mining is two fold. First, sensitive raw data like identifiers, names, addresses, should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy.

II. Background

The literature review is done to get an insight of the basics of privacy preserving data mining [1, 2, 3The basis of the literature review is to critically establish the extent and depth of existing techniques to preserve sensitive data and knowledge from large database, and to find efficient approach for preserving private or sensitive data

A. Method of K- Anonymity

When releasing micro data for research purposes, one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Sweeney introduced the *k*-anonymity privacy requirement, which requires each record in an anonymzed table to be indistinguishable with at least *k*-1 other records within the dataset, with respect to a set of quasi-identifier attributes [4]. To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization. Anonymization [5, 6] is a process that removes or replaces identity information from a communication or record. Table I shows micro data. TableII represents voter's registration data. TableIII shows an example of 2-anonymous generalization for TableI. Even with the voter registration list, an adversary can only infer that Jim may be the person involved in the first 2 tuples of TableIII, or equivalently, the real disease of Jim is discovered only with probability 50%. While *k*-anonymity protects against identity disclosure, it does not provide

DOI: 01.IJRTET.8.1.58

TABLE I.MICRODATA

ID	Age	Gender	Zipcode	Disease
1	26	M	83661	Headache
2	24	M	83634	Headache
3	31	M	83967	Toothache
4	39	F	83949	Cough

TABLE II. VOTER REGISTRATION LIST

ID	Name	Age	Gender	Zipcode
1	Jim	26	M	83661
2	Jay	24	M	83634
3	Tom	31	M	83967
4	Lily	39	F	83949

TABLE III. A-2 ANONYMOUS TABLE

ID	Age	Gender	Zipcode	Disease
1	2*	M	836**	Headache
2	2*	M	836**	Headache
3	3*	*	839**	Toothache
4	3*	*	839**	Cough

sufficient protection against attribute disclosure because the of the *k*-anonymity model stem from the two assumptions [7, 8].This method has drawback of homogeneity and background attack. A linking attack is executed by taking external tables containing the identities of individuals, and some or all of the public attributes [9].

B. Random Perturbation

This method [10], can deal with character type, Boolean type, classification type and number types of discrete data, and to facilitate conversion of data sets, it is necessary to preprocess the original data set. The data preprocessing is divided into discrete data, attribute coding, data sets coded data set, three parts, this paper uses the method of average region to disperse the continuous data. Discrete formula is as follows: A (max) - A (mi n)/n = length. A is continuous attributes, n is the number of discrete, length is the length of the discrete interval. When the interval length is a decimal, round to the nearest integer, the first interval of discrete begin from A (min), the last interval is A (max). The method [11, 12, 13], does not reconstruct the original data values, but only reconstruct distribution.

C. Blocking based Method

Blocking technique applies to applications where we can store unknown values for some attributes, when actual values are not available or confidential [1] .This method replaces the 1's or 0's by unknowns ("?") in selected transactions. So, that rule will not be generated from the dataset. The goal of the algorithm presented here is to obscure a given set of sensitive rule by replacing known values with unknown values. For each sensitive rule, it scans the original database and find outs the transactions supporting sensitive rules. We can say transaction supports any rule when the left side of the rule pair is a subset of attribute values pair of the transaction and the right hand side of the rule is same as the class attribute of the transaction. Then for each transaction that supports sensitive rule, algorithm places "?" (Unknown) values in place of attribute value which appears in rule. This procedure continues until all the sensitive rules are hidden. Finally the sanitized dataset which contains unknown values is released to public [14].This method is easy to implement but gives information loss.

D. Cryptographic Technique

B. Pinka s introduced cryptographic technique [15], is popular mainly for two reasons. Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining. This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records [16].

E. Condensation Approach

Charu C. Aggarwal and Philip introduced Condensation approach which constructs constrained clusters in the data set, and then generates pseudo-data from the statistics of these clusters [17, 18]. This approach uses a methodology which condenses the data into multiple groups of predefined size, for each group, certain statistics are maintained. Each group has a size at least k, which is referred to as the level of that privacy-preserving approach. It gives information loss.

III. PROBLEM DEFINITION

Given a dataset T with k attributes A1, A2…..AK. In order to satisfy the requirement of data privacy preserving, first attribute transitional probability matrix is used to dataset T, and generate the private dataset D. Next K-anonymity method is applied on private dataset D. It is noticed that we cannot precisely find out the records of T from D based on data randomization and generalization.

IV. PROPOSED METHOD

Through the study of literature survey has identified different privacy preserving techniques still there is a need to find the efficient method to preserve privacy in large database. The proposed Hybrid algorithm has two main advantages. The proposed method protects private data with no loss of information which makes usability of data, and also data is reconstructed.

A. Hybrid Algorithm

Inputs: Original training dataset T, Transition probability matrix P, Mapping Matrix M with size 1*j between T and P

Output: Conversed training dataset D, Derived table.

Method:

a)Select the attribute from table T.

b)Generate probability matrix P randomly with size j*j.

c)Generate Mapping matrix M randomly

DOI: 01.IJRTET.8.1.58

d) Assign each P (P1, P2… Pj) to the column of T (T1, T2… Tj) randomly by M.

e) Rearrange element of T with respect to highest value of P location. If P location is already used, go for next highest location, if value of the P of two or more location is same then choose the left hand side value.

f) Recombine T matrix.

g) Re-substitute in table.

h) Apply K-anonymity on table based on selected attribute.

i) Generalize the value, if numeric takes the range of lowest and highest value.

j) Stop.

The proposed algorithm first selects attribute or quasi identifier from table T. After that it generates probability matrix P and mapping matrix M randomly. Element of T are rearranged with respect to highest value of P location. If P location is already used, then it will go for next highest location, if value of the P of two or more location is same then it will choose the left hand side value. All values are re-substituted in table T. After that numeric values are generalized by taking the range of lowest and highest value.

B. An Example

We have provided an example in tableIV to further explain the application of proposed algorithm. TableIV exemplifies medical data to be released. Names are removed in the medical data in order to hide the identities of the individuals. Age, gender, zip code are quasi identifier and disease is sensitive attribute. In tableIV, we considered T1=Age values, T2= Gender, and T3= Zip Code. We randomly generate 3 matrices with the size 7*7 (P1, P2, P3) respectively because column size is 7.We have used patient data set here.

TABLE IV. PATIENT DATASET

Sr. no	*Age*	*Gender*	*Zip code*	*Disease*
1	33	M	600018	Flu
2	29	F	600008	Stomach cancer
3	21	M	600006	Bronchitis
4	31	M	600009	Gastritis
5	22	M	600006	Bronchitis
6	60	M	600019	Flu
7	25	F	600006	Bronchitis

We have generated random matrix in java.

P1=
0.78	0.69	0.38	0.49	0.39	0.56	0.8
0.9	0.56	0.25	0.56	0.78	0.37	0.9
0.12	0.89	0.31	0.57	0.59	0.89	0.7
0.15	0.31	0.51	0.75	0.69	0.9	0.3
0.35	0.51	0.37	0.36	0.70	0.3	0.8
0.8	0.46	0.39	0.39	0.69	0.7	0.9
0.7	0.26	0.49	0.69	0.50	0.8	0.6

P2=
0.8	0.9	0.6	0.5	0.7	0.9	0.1
0.8	0.3	0.6	0.7	0.9	0.3	0.5
0.3	0.6	0.3	0.2	0.1	0.3	0.1
0.5	0.7	0.8	0.9	0.3	0.1	0.1
0.2	0.3	0.6	0.7	0.9	0.5	0.3
0.3	0.5	0.8	0.9	0.3	0.9	0.1
0.3	0.8	0.9	0.1	0.3	0.3	0.6

P3=
0.8	0.8	0.3	0.5	0.2	0.3	0.3
0.9	0.3	0.6	0.7	0.3	0.5	0.8
0.6	0.6	0.3	0.8	0.6	0.8	0.9
0.5	0.7	0.2	0.9	0.7	0.9	0.1
0.7	0.9	0.1	0.3	0.9	0.3	0.3
0.9	0.3	0.3	0.1	0.5	0.9	0.3
0.1	0.5	0.1	0.1	0.3	0.1	0.6

After generating probability matrix P, we have generated mapping matrix randomly M = [2, 3, 1]. It means that we have matched P2 to T1, P3 to T2 and P1 to T3. In P2, positions of largest value of each line are 2, 5, 1, 4, 3, 6, and 7 .So successively choose the 2nd (29), the 5th (22), the 1st (33), the 4th (31), the 3rd (21), the 6th (60) and the the 7th (25) value of T1 to form D1. Similarly we can get D2 and D3. Table V shows values after applying randomization method on tableIV patient dataset. Final derived table is shown in tableVI.

TABLE V. CONVERTED TABLE D

Sr. no	*Age*	*Gender*	*Zip code*	*Disease*
1	29	M	600006	Flu
2	22	F	600018	Stomach cancer
3	33	M	600008	Bronchitis
4	31	M	600019	Gastritis
5	21	F	600006	Bronchitis
6	60	M	600006	Flu
7	25	M	600009	Bronchitis

TABLE VI. FINAL DERIVED TABLE

Sr. no	*Age*	*Gender*	*Zip code*	*Disease*
1	29-60	M	600***	Flu
2	20-29	F	600***	Stomach cancer
3	21-33	M	600***	Bronchitis
4	21-33	F	600***	Bronchitis
5	30-39	M	600***	Gastritis

V. EXPERIMENTAL SETUP AND RESULTS

We have taken dataset from [19, 20, 21, 22]. The proposed method is applied on medical dataset containing 1000 records with attributes ID, age, sex, zip code and disease. We have considered disease as a sensitive attribute. We have implemented algorithm in JDK 1.7 and Net beans 7.1.2. and made to run on Intel Dual core, 2.0 GHz, 2GB RAM. In figure 1 the proposed method preserve better privacy as compared with existing k-anonymity and block based method. Figure1shows comparison of performance with other method. Figure2 shows comparison of proposed method for data reconstruction.

CONCLUSION

Privacy preserving is growing field of research. From the literature review, it is clear that Anonymity technique has drawback of homogeneity and background attack. Random perturbation technique does not provide usability of data. Blocking method gives information loss. In the proposed method as we combined K-anonymity with randomization technique it is difficult for attacker to identify homogeneity and background attack. Also it protects private data with better accuracy and gives no loss of information which makes

DOI: 01.IJRTET.8.1.58

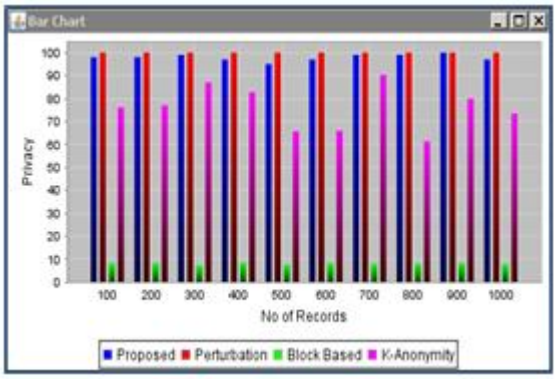

Fig. 1. Analysis of privacy preserving algorithms

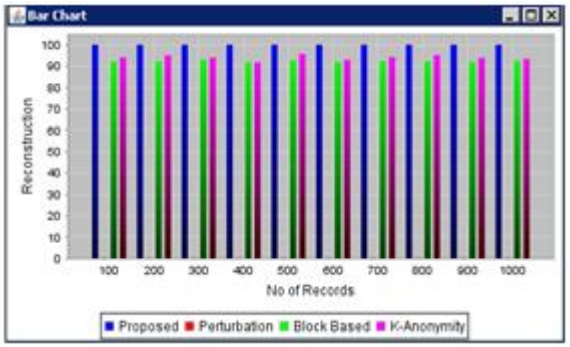

Fig. 2. Comparison of data reconstruction

usability of data, and also data can be reconstructed.

REFERENCES

[1] Savita Lohiya, Lata Ragha ,"Privacy preserving in Data Mining using Hybrid Approach", IEEE International Conference on Computational Intelligence and Communication Network (CICN) ,2012.

[2] Anita A. Parmar, Udai Pratap Rao, "Blocking Based approach for Classification Rule Hiding to Preserve the Privacy in Database", International Symposium on Computer Science and Society (ISCCS) , pp.323-326, 2011

[3] Jian Wang, Yongcheng Lou, Yen Zhao, Jiajin Le, "A Survey on Privacy Preserving Data Mining", International Workshop on Database Technology and Applications, pp.111-114, 2009.

[4] L. Sweeney, "K-anonymity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570, 2002.

[5] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation", IEEETrans. Knowledge and Data Engg., vol. 19, no. 5, pp. 711-725, May 2007.

[6] S.Vijayarani, A.Tamilarasi, M.Sampoorna, "Analysis of Privacy Preserving K Anonymity Methods and Techniques", Proceedings of the International Conference on communication and Computational Intelligence, pp.540-545, December 2010.

[7] Yan Zhu, Lin Peng, "Study on K-anonymity Models of Sharing Medical Information", International Conference on Service Systems and Service Management, pp.1-8, 2007.

[8] Weijia Yang,"Knowledge Reserving in Privacy Preserving Data Mining", Second international Symposium on Intelligent Information Technology Application, pp.855-859, 2008.

[9] E. Poovammal , M. Ponnavaikko , "Task Independent Privacy Preserving Data Mining on Medical Dataset", International Conference on Advances in Computing, Control and Telecommunication Technologies ,pp.815-818, 2009.

[10] Xiaolin Zhang, Hongjing Bi, "Research on Privacy Preserving Classification Data Mining Based on Random Perturbation", International Conference on Information Networking and Automation (ICINA), pp.173-178, 2010.

[11] H .Kargupta, S. Datta, Q. Wang, and K. Siva Kumar, "On the privacy preserving Properties of Random Data Perturbation techniques", Proceeding of International Conference on Data Mining, pp. 99 - 106 , 2003.

[12] K. Chen and L. Liu, "Privacy Preserving Data Classification with Rotation Perturbation", Proceedings of the Fifth International Conference of Data Mining (ICDM'05), pp.589-582, 2005.

[13] Li Liu; Bhavani Thuraisingham, "The Applicability of the Perturbation Model-based Privacy Preservig Data Mining for Real-world Data", Sixth IEEE International Conference on Data Mining Workshops, pp.507-512, 2006.

[14] Jinfei Liu, Jun Luo, and Joshua Zhexue Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity Requirements", International conference on Data Mining Workshops, pp.666-670, 2011.

[15] Vassilios S. Verykios, Elisa Bertino, "State-of-the-art in Privacy Preserving Data Mining", Proceeding of Special Interest Group on Management of Data (SIGMOD) Record, Vol. 33, No. 1,pp.50-57, 2004.

[16] Y. Lindell , B.Pinkas, "Privacy preserving data mining", Journal of Cryptology, 5(3), 2000.

[17] Haisheng Li , "Study of Privacy Preserving Data Mining", Third International Symposium on Intelligent Information Technology and Security Informatics, pp.700-703, 2010.

[18] Charu C. Aggarwal , Philip S. Yu, "A condensation approach to privacy preserving data mining", International Conference on Extending Database Technology (EDBT), pp. 183–199, 2004.

[19] Official website of IBM Research software group, http://www-958.ibm.com/datasets/datasetof-diabeticpatient.

[20] Officia website of spss online training workshop, http:// people.cst.cmich.edu/lee1c/spss/Prj_cancer_data.htm

[21]website of Keng Ridge Bio-medical Dataset, http://datam.i2r.a-star.edu.sg/datasets/krbd/

[22] UCI Repository of Machine Learning Databases, www .ics. uci. edu/~ mlearn/MLRepository.html

DOI: 01.IJRTET.8.1.58